A Contribution to Deep Learning Approaches for Automatic Classification of Volcano-Seismic Events: Deep Gaussian Processes

Miguel López-Pérez, Luz García, Carmen Benítez, and Rafael Molina, *Senior Member, IEEE*

***Abstract*— The automatic classification of volcano-seismic events is a key problem in volcanology. Due to its complexity, deep learning (DL) techniques have become the tool of choice for this problem, outperforming classical classifiers. The main drawback of this approach, when applied to the classification of volcano-seismic events, is its tendency to overfit because of the small-size available databases. In this work, we propose and analyze the use of the Gaussian processes (GPs) and Deep GPs (DGPs), and their hierarchical extension, for volcano-seismic event classification. We empirically prove the adequacy of the proposed modeling with an insightful and exhaustive comparison with state-of-the-art DL-based methods on a seismic database recorded at "Volcán de Fuego," Colima, Mexico. The hierarchical structure of DGPs and the reduced number of parameters to be automatically estimated become essential to achieve excellent performance even on small databases, capturing well the complex patterns of seismic signals for all classes and, in particular, for those that have been hardly observed.**

***Index Terms*— Deep Gaussian processes (DGPs), deep learning (DL), Gaussian processes (GPs), geophysical signal processing, geoscience and remote sensing, remote monitoring, remote sensing, signal processing, volcanic activity, volcanoes.**

I. Introduction

GEOPHYSICAL processes, such as displacements of magma and other fluids or gases, or fractures of solid materials in volcanic areas, are derived from the exchange of elastic energy between volcanic structures and their surroundings. Seismic signals registered by stations deployed near volcanoes capture elastic waves that reflect such exchanges. Their study provides very valuable information. When properly interpreted, seismic signals offer a useful insight into the internal dynamics of the volcano. Source mechanisms originating them can be inferred from their analysis, together with information about the Earth's crust materials traversed during the trip of the elastic wave toward the station registering it [1], [2].

Detection and classification of seismic events consist of processing seismic registers to spot events and associate them with their original geophysical source mechanism based on the characteristics of the signal. The source mechanism inference is a complex task given the number of additional factors that influence the signal arriving at the seismic station. The degree of elasticity/anisotropy of materials in the source location, distance to the station, characteristics of the propagation path, or frequency response of the registering instrument are examples of them. Once detected, the spotted patterns in the sequence of events are analyzed to understand the physical model explaining the dynamics of the volcano. They are also used in applications, such as early warning monitoring systems based on the detection of events precursors of eruptions.

In the last decades, the amount of seismic data available has increased enormously together with the computing and storage capacity. These facts have encouraged the geophysical community to explore the use of machine learning (ML) algorithms for automatic classification of seismic events [3]–[5]. ML techniques avoid the tedious and repetitive work of manual labeling, often done by geophysical experts, and increase the capacity to process enormous volumes of data. They capture complex data correlations not detectable by human experts. There is a wide range of possible ML algorithms usable for automatic classification of seismic events. The election of the approach depends on factors, such as the dimensionality of the data and corresponding classes, the size of the labeled training database, the continuous/isolated classification objective needed, or the interpretability of the model searched for.

Within the field of seismicity, the classification of volcano-seismic events presents specific challenges derived from their origins. Simultaneous seismic events related to liquid and/or gas-solid processes take place in the volcanic scenario. Tremors, long-period events (LPEs), or surface effects, such as rockfalls, landslides, or pyroclastic density flows, might happen simultaneously generating complex seismic registers with overlapped events. In addition, volcanic regions present changing propagation and site properties. Sismo-volcanic

Manuscript received January 28, 2020; revised July 15, 2020; accepted August 27, 2020. This work was supported in part by the Ministerio de Economía y Competitividad (MINECO) under Project DPI2016-77869 and Project A-TIC-215-UGR18; in part by the Ministerio de Ciencia e Innovación under Project PID2019-105142RB-C22 and Project PID2019-106260GB-I00; and in part by the European Commission (Action MSCA-ETN-ITN) under Project SEP-210560672. *(Corresponding author: Miguel López-Pérez.)*

Miguel López-Pérez and Rafael Molina are with the Department of Computer Science and Artificial Intelligence, University of Granada, 18010 Granada, Spain (e-mail: mlopez@decsai.ugr.es).

Luz García and Carmen Benítez are with the Department of Signal Theory, Telematics and Communications, University of Granada, 18071 Granada, Spain.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2020.3022995

sources are often shallower compared with tectonic ones. As a consequence, near-source and surface-propagation effects complicate the analysis of the seismic signal. The labeling task must, therefore, be carried out by expert geophysicists with a deep knowledge of the particular volcano generating the data. This is a difficult, tedious, and time-consuming task that requires deep expert knowledge and strict maintenance of the labeling criteria. For all these reasons, large enough databases with high-quality labels are scarce but extremely necessary to improve the knowledge of the volcanic structures and predict their behavior.

Supervised ML techniques for automatic classification of isolated volcano-seismic events started around 2005 with the usage of artificial neural networks (ANNs) in the pioneer [6]. Since then, interesting applications, such as in [7], models based on support vector machines (SVMs) [8], combination of several shallow classifiers, such as ANN and SVM [9] or ANN, and genetic algorithms [10], have been developed. In parallel, the hidden Markov models [11]–[14] have been introduced to model temporal structures, providing approaches to successfully detect and classify events in continuous seismic registers.

The deep learning (DL) approaches, with a higher degree of abstraction and knowledge extraction for complicated data sets, became popular after the proposals of Hinton *et al.* [15] in 2006 and Bengio *et al.* [16] in 2012 (accompanied by important advances in computational power). These two works proposed, respectively, to use the restricted Boltzmann machines (RBMs) and denoising autoencoders (DAs) to initialize hidden layers via unsupervised layer-by-layer training, proving that Deep Networks could be trained well, with more optimal initializations and useful learned representations of the data.

DL was first applied to image processing and speech and has spread its usage to many disciplines, with attractive applications in the field of seismology. Examples of them are the automatic P-phase picking approach in [17], the skip connection CNN proposed in [18] to detect geyser related events in continuous registers, or the usage of deep convolutional autoencoders for seismic signal clustering in [19]. Classification of volcano-seismic signals using deep neural networks (DNNs) was first presented by Titos *et al.* [20] with the implementation of a deep belief network (DBN) and a stacked denoising autoencoder (sDNA). Their classification performance was compared with the state-of-the-art isolated events classifiers on the seismic events database of the Volcán de Fuego de Colima, México. The work in [21] implemented and compared three recurrent neural network (RNN) architectures (Vanilla, LSTM, and GRU) to detect and classify volcano-seismic events from the Volcán de Decepción, Antarctica, in continuous registers. Unfortunately, the use of DL techniques is based on the availability of large amounts of data. To overcome the lack of large databases of labeled volcano-seismic events necessary for effective classification with DL architectures, transfer learning approaches based on DL have been explored in [22] and [23].

In the ML scenario described so far, GP models were introduced in 2006 [24]. They are nonparametric probabilistic models that deal with uncertainty in prediction and modeling. Interesting connections between DNNs and GPs were studied in [25], where the correspondence between GPs and priors for *infinitely wide* DNNs was established. Their expressiveness and robustness to overfit have been largely praised. The prior information in the kernel function of the GPs acts like a regularizer, making them suitable for not very large databases, which is the case in volcanology. This is in contrast to neural networks that have to learn a huge amount of parameters to estimate a complex model, and so, they tend to overfit on small databases. Furthermore, as Lawrence [26] indicates in his post, *the next generation of data-efficient learning approaches relies on us developing new algorithms that can propagate stochasticity or uncertainty right through the model* (see [27] and the seminar thesis [28]).

Although GPs are very flexible, they suffer from a severe limitation. They are commonly used with stationary kernels, which makes them unsuitable for complex patterns, e.g., functions that combine flat regions with high-variability ones. Recent advances have shown that any number of GP models can be stacked to implement deep hierarchies. These hierarchical models maintain the main advantages of GPs while learning more abstract and complex models. Deep Gaussian processes (DGPs) were first introduced in [29] in 2013, and their probabilistic DL modeling was very promising but the inference procedure complicated. In 2017, Salimbeni and Deisenroth [30] introduced the doubly stochastic variational inference model for DGPs that, since then, became the current state of the art for DGP inference.

GPs, but not DGPs, have been used for different tasks in seismic problems although none of them have been ever used before for automatic seismic-event classification. Especially, GPs have been used for regression in seismic problems with promising results in this field. Moore and Russell [31] proposed a generative model for seismic monitoring. This model can recover weak events from the raw signal. They used GPs over wavelet parameters to predict detailed waveform fluctuations based on historical events while degrading smoothly to simple parametric envelopes in regions with no historical seismicity. Moore and Russell [32] proposed a new approximation for large-scale GPs, especially for GP latent variable models (GPLVMs). They proposed to approximate the marginal likelihood of the full GP via a random Markov field in which local GPs are connected by pairwise potentials. This approximation allows us to efficiently perform inference for spatial data, and it was applied successfully to seismic location. Noori *et al.* [33] used GP regression for anomaly detection, more specifically, for fault detection in seismic data. Since the used GPs expected smooth functions, their results show that fault points can be detected when the smooth trend of layers is disrupted by faulting.

This article represents, to the best of our knowledge, the first contribution to the use of GPs and DGPs for automatic seismic-event classification. An approach that is tested here on the seismic data set recorded at the Volcán de Fuego de Colima, Colima, Mexico. Due to the complex character of this classification problem, the current state-of-the-art methods are based on hierarchical deep models. We show here that

GPs outperform all the shallow classifiers and that they are competitive to DNNs. The experiments also show that the two-layer DGP model outperforms DNNs, in particular in classes hardly represented. Additional experiments indicate that GPs and DGPs can learn good models even when the database is small. When data are scarce, GPs are the best-performing models. With more data, deeper models, such as the four-layer DGPs, provide better results. The study on the prediction confidence of each model shows that GP-based methods obtained probabilities closer to 1 than DNNs.

The rest of this article is organized as follows. In Section II, we provide a brief introduction to both GPs and DGPs. This introduction is expanded in the Appendix where a complete theoretical and intuitive description of GPs and DGPs for multiclass classification problems is included. In Section III, we carry out an insightful and exhaustive experimental analysis whose goal is to compare GPs and DGPs to current state of the art both shallow and deep classifiers on the database recorded at Volcán de Fuego de Colima [20]. Conclusions are drawn in Section IV.

II. Deep Gaussian Process Classifier

In this section, we provide a brief introduction to the use of GPs and DGPs for multiclass classification problems. An extended and more detailed introduction can be found in the Appendix.

A multiclass classification problem with K classes consists of N labeled instances $\{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$, where $\mathbf{x}_n \in \mathbb{R}^D$ is the feature vector and $y_n \in \{1, \ldots, K\}$ is the class label of the nth instance. For each instance $\mathbf{x}_n$, its label y_n is modeled using K latent variables $\mathbf{f}_{n,:} = \{f_k(\mathbf{x}_n)\}_{k=1}^{K}$ through a specific likelihood $\mathrm{p}(\mathbf{y}|\mathbf{f}_{n,:})$. In this work, we utilize the robust max likelihood, which prevents overfitting in GPs.

In a GP-based formulation of a supervised problem, we assume that the distribution of $\mathbf{f} = (f_1, \ldots, f_n)^{\mathrm{T}}$ given $\mathbf{X}$ is a multivariate normal, where we assume zero mean for simplicity, and a kernel function $k(\cdot, \cdot)$ defines our covariance matrix. In this article, we use the squared exponential (SE) kernel defined as $k_{\mathrm{SE}}(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 \exp((-||\mathbf{x}_i - \mathbf{x}_j||^2/2l^2))$, where the parameters σ and l will be estimated from the observations (see Figs. 8, 9, and 11 for a better understanding of GPs). For scalability, we also defined $M \ll N$, inducing points $\mathbf{u}_m$, which are the realization of the GP in the locations $\mathbf{z}_m$, that is, $\mathbf{u}_m = \mathbf{f}(\mathbf{z}_m)$. The inducing points summarize information from the entire data set in a few points. Their locations are learned in the optimization process too. The posterior distribution for this model is not tractable so an approximate inference method has to be used. In this work, we follow the scalable variational inference for GPs (SVGP) [34]. From here on, and to make explicit the inference procedure used, we will refer to the single-layer GP as SVGP.

SVGPs are flexible nonparametric probabilistic models very frequently used in classification and regression problems. However, these models can only represent a restricted class of functions. To overcome this limitation, hierarchical models based on GPs were proposed [29]. DGPs use the outputs of a standard SVGP as the input to another SVGP. If this is repeated L times, we obtain a hierarchy of SVGPs, which is known as a DGP with $L + 1$ layers. Due to its hierarchical structure, it achieves a greater level of abstraction and can capture more complex patterns (see Figs. 13 and 14 for a better understanding of how DGPs tackle the complexity in a toy example). Volcano seismic signals are very complex with classes that are difficult to distinguish. We will see that the DGPs are very suitable for our classification task.

III. Practical Application: Automatic Events Classification for the 'Volcán De Fuego De Colima', México

A. Database Description

Section III analyses the performance of SVGP and DGP methods for the classification of volcano-seismic events. Classification experiments are carried out using a database of 9.332 seismic events registered at the Volcán de Fuego de Colima [35]. These registers and their labels are the result of a careful and demanding process of expert analysis and review to eliminate human artifacts and noise, to identify source mechanisms, and to analyze how site and path effects can influence waveforms. The labeled database contains seven different events (classes) with diverse spectral and temporal characteristics, associated with seven corresponding source mechanisms (REG, VTE, LPE, TRE, EXP, COL, and NOISE). They can be grouped as follows.

1) *Events Originated by Fractures of Solid Materials in the Earth's Crust:* Regional earthquakes (REG) and volcano-tectónic earthquakes (VTE). As a result of the fracture, elastic waves containing P- and S-wave components associated, respectively, to longitudinal and shear displacements are generated. If the fracture occurs in the surrounding of the volcano, the event associated is identified as "VTE" and contains high frequencies reaching up to 40 Hz with durations from a few to tens of seconds. On the other hand, fractures that might occur in fault planes beyond the volcanic region can be registered by the seismometers in the volcanic area, being labeled as "REG." REG events contain frequencies lower than those of VTEs because the higher ones have been absorbed through the propagation path from the fracture source location to the registering station. The database used in this work contains 1.738 VTEs and 455 REGs.
2) When no fracture occurs, volumetric modes of deformation of the volcanic structure (often triggered by displacements of water, gas, or magma), produce LPEs. They show frequencies of a few Hz (between 1 and 6 Hz for the Volcan de Colima) and durations of a few seconds. Having spectral characteristics and source mechanisms similar to those of LPEs but much longer duration, Volcanic Tremors (TRE) identify a series of harmonic signals with sustained amplitude and variable duration from minutes to hours. The database used in this work contains 2.699 LPEs and 1.170 TREs.
3) There are also certain events associated with the external activity of the volcano. Often, sudden emissions of gas

and ash to the atmosphere occur and are recorded by seismometers, receiving the name of Explosions (EXP). They are characterized by a short-duration LPE, followed by high-frequency signals with a narrow energy peak that can reach up to 20 Hz. Surface lava movements, or Lava Flows, (COL) with durations of minutes and frequencies between 5 and 10 Hz are also associated with the external dynamics of the volcano. The database used in this work contains 2.699 LPEs, 278 EXPs, and 1.406 COLs.

4) Finally, seismic noise (NOISE) registered by stations in the absence of volcanic source mechanisms presents diverse amplitudes, frequencies, and durations depending on its nature (wind, sea, rain, cultural noise, and so on). The database used contains 1.586 NOISE examples.

For comparison purposes, we follow the approach in [20]. The events used to feed the models are parameterized to create input feature vectors with 21 features. Seismic registers are first filtered in the band 1–25 Hz. Then, regardless of their duration, they are divided into three segments of equal length (beginning, central part, and ending of the event). After that, following a common parameterization in the field, for each segment, a feature vector of five linear predictive coding (LPC) coefficients is calculated. The 15-feature vector so built is completed with six statistical features proposed in [36]. Features 16–18 parameterize the impulsiveness of the signal in the time domain by calculating the 20th, 50th, and 80th cumulative-sum percentiles of the signal's amplitude. Following the same approximation in the frequency domain, features 19–21 calculate the 20th, 50th, and 80th cumulative-sum percentiles of the signal's power spectral density.

B. Experiments Description

The chosen methods for this study are the following: the single-layer SVGP (SVGP) and the two-layer (DGP2), three-layer (DGP3), and four-layer (DGP4) DGPs. We also include an exhaustive and insightful comparison to state-of-the-art shallow and deep classifiers. The selected shallow classifiers are SVM with linear (SVM-Lin) and radial (SVM-Rad) kernels, random forest (RF), and a single-layer MLP. The deep classifiers are the following DNNs: DBN with two (DBN-H2) and three (DBN-H3) hidden layers and sDNA with two (sDA-H2) and three (sDA-H3) hidden layers. Configuration details for these classifiers, which were tuned performing grid searches for the optimal number of neurons per layer, are fully described in detail in [20].

The data set is carefully split into four folds to perform a fourfold cross-validation analysis. Taking into account the unbalanced nature of the classes of volcano-seismic events, folds are carefully checked to ensure well-balanced statistically representative experiments. For the sake of comparison between GPs and other DL approaches, the same database and folds used in [20] are used in the present experiments. Given the need to tune the system architecture, on each round of the cross-validation, two folds are used for training: one to search the optimal configuration (grid search for the possible numbers of neurons per layer) and another to evaluate the classification results.

We use three different metrics to assess the performance: the f1 score, accuracy, and log loss. We define the multiclass accuracy and log loss as

$$\text{accuracy} = \frac{\text{No. events classified correctly}}{\text{Total no. events}} \tag{1}$$

$$\text{log loss} = \frac{-1}{N}\sum_{n=1}^{N}\log(\text{p}(y_n)\cdot \mathbf{e}_{k_n}) \tag{2}$$

where N is the total number of events, the dot $(\cdot)$ denotes the scalar product, and $\mathbf{e}_{n_k}$ is the one-hot encoding vector of the true class of the nth instance. Notice that the accuracy is the percentage of global success, and the log loss measures not only the success but the confidence of the classifier. We define the f1 score per class using the true positives (TPs), false negatives (FNs), and false positives (FPs) as

$$\text{f1 score} = \frac{2\times \text{TP}}{2\times \text{TP} + \text{FN} + \text{FP}} \tag{3}$$

and then, we take the average of them to obtain the multiclass macro-average f1 score. Notice that this metric penalizes the misclassification of samples coming from underrepresented classes while the log loss and accuracy do not.

To provide a deep insight into the particular needs in the classification of volcano-seismic events, the rest of the experimental section has been structured as follows. First, in Section III-C, we study the behavior and selection of hyperparameters using the validation set. In addition to the selection of the model configuration, this experiment also provides a better understanding of the presented models. Then, in Sections III-D and III-E, we assess the generalization capability of the models on the test set. Finally, additional experiments of relevant interest in the area of knowledge are reported. Given the lack of large-high-quality labeled databases, the robustness of the classification against different sizes of the data set is studied in Section III-F. In addition, in order to handle the difficulties to classify some events that could correspond to diverse source mechanisms (including overlapped ones), confidence measures of the predictions for the different classifiers are studied in Section III-G.

C. Selection of SVGP and DGP Hyperparameters

In contrast to other classifiers where an exhaustive grid search is used for hyperparameter tuning, in GP-based methods, almost all the parameters are estimated automatically and learned through an optimization process. Following common practice [30], we utilize the same number of hidden units in each layer. Since, in this problem, we have a reduced number of features, we set it to 7 after an empirical search. We use the SE kernel defined in (8). In this model the length scale l associated with all the features is the same. This is very useful to avoid overfitting in scenarios with small databases, but features frequently have different discriminative power. In the experiments, for an exhaustive comparison, we also use the automatic relevance determination (ARD) model, whose

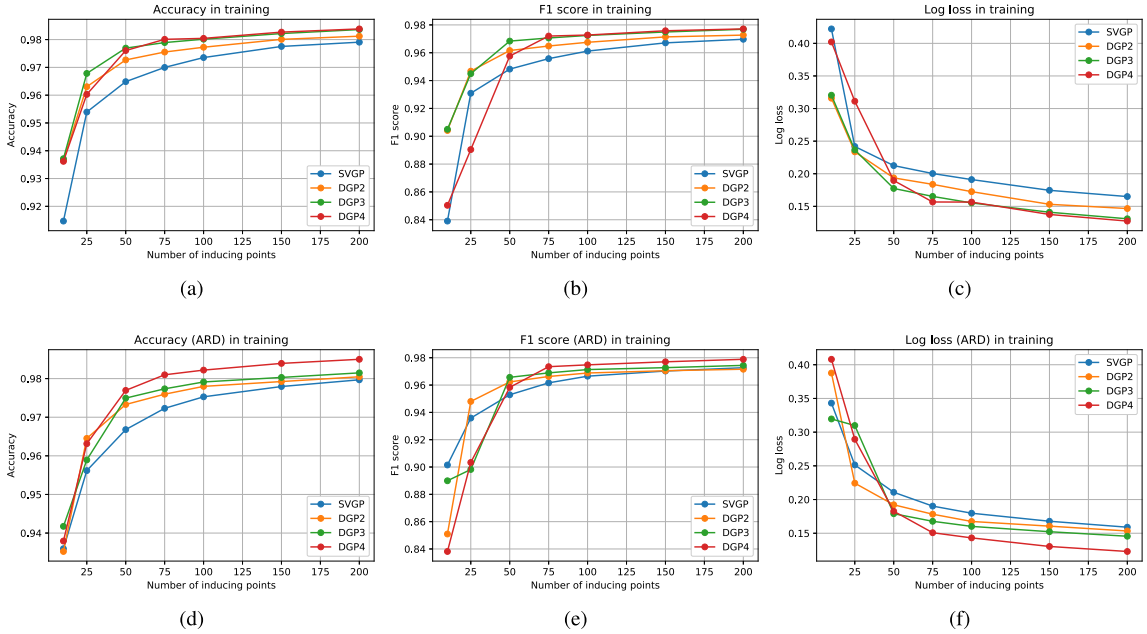


Fig. 1. Results on the training set for GP-based models. Using (a)–(c) SE kernel and (d)–(f) SE-ARD one. Each column corresponds to a different metric (from left to right): accuracy, f1 score, and log loss.

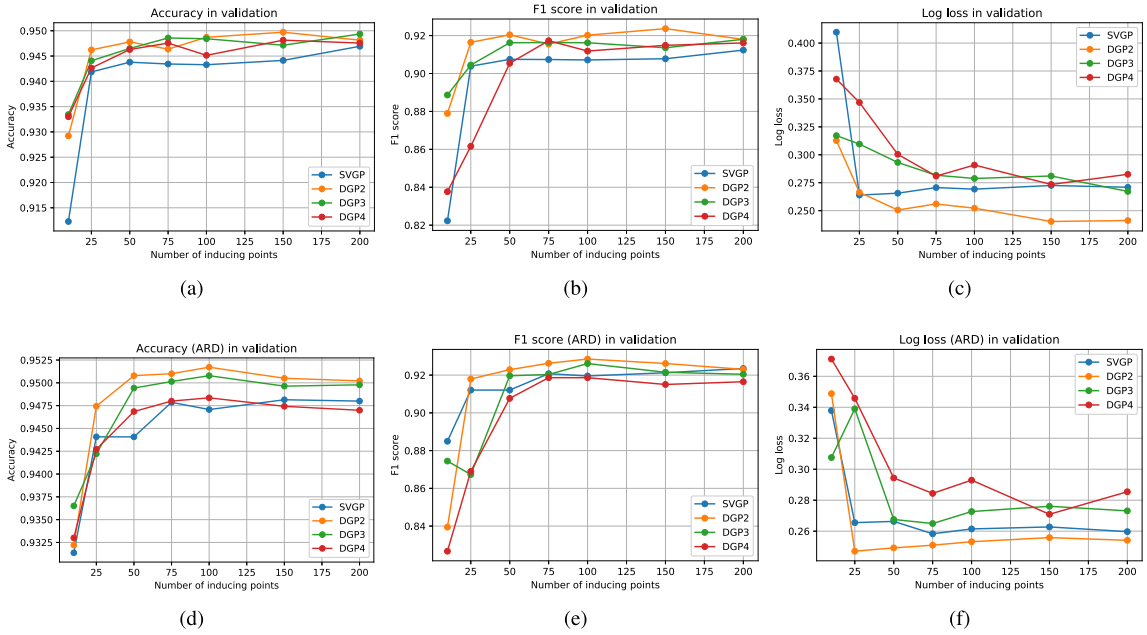


Fig. 2. Results on the validation set for GP-based models. Using (a)–(c) SE kernel and (d)–(f) SE-ARD one. Each column corresponds to a different metric (from left to right): accuracy, f1 score, and log loss.

kernel is defined by

$$k_{\text{SE-ARD}}(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 \exp\left(-\sum_{d=1}^{D} \frac{||x_i(d) - x_j(d)||^2}{2l_d^2}\right) \quad (4)$$

where $x_i(d)$ is the dth coordinate of the feature vector $\mathbf{x}_i$, and l_d is its length-scale parameter.

To adjust the number of inducing points, we choose the following grid analysis: 10, 25, 50, 75, 100, 150, and 200 points. We report the following metrics for every combination of inducing points and kernel (SE or SE-ARD): accuracy, f1-score and log loss, for both training and validation sets.

Results on the training set are shown in Fig. 1. As the number of layers and inducing points increases, the models perform better except for a slightly noisy behavior when few inducing points are used. On the validation set (see Fig. 2), more complex architectures do not lead to better models. Once we have a sufficiently high number of inducing points, we cannot capture more information to have a better performance in the validation set. To summarize

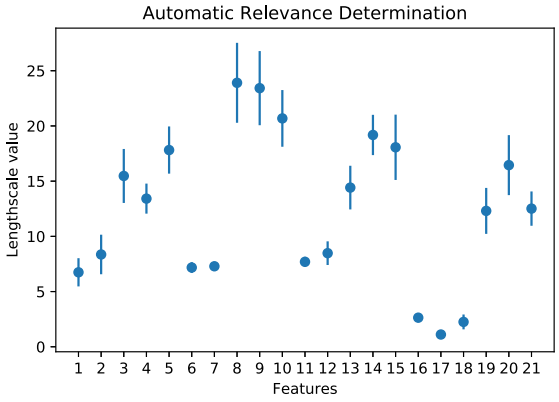


Fig. 3. Estimated length-scale values for the SVGP in test. The points represent the averaged values, and the bars represent the standard deviation. Lower (downward) represents more importance of that feature for the classifier.

the essential information of the analysis, 100 points are enough.

We also compare both the SE and SE-ARD kernels to find out whether there are features more relevant than others. We can see that the ARD results are, in general, slightly better. For example, looking at Fig. 2, SVGP, DGP2, and DGP3 reach a 0.92 f1 score value with ARD, while SE DGP2 hardly reaches this value. Such improvement will be helpful when detecting underrepresented classes since this global metric gives weight to them.

To conclude this section, in general, the SE-ARD kernel model performs better than SE. Furthermore, SE and SE-ARD become stable once they reach 100 inducing points. Based on these results, we chose the SE-ARD kernel and 100 inducing points as hyperparameters of the SVGP and DGP models for the subsequent evaluation of the test sets.

D. Performance of Shallow Classifiers

In this section, we assess the generalization capability of shallow SVGP with 100 inducing points and SE-ARD kernel against the shallow state-of-art classifiers: SVM with linear (SVM-Lin) and SE (SVM-SE) kernels, RF with 120 estimators, and a single-layer MLP.

In Table I, we report per class and global f1 scores and accuracy. In addition, we calculated the 95% confidence interval for the global metrics. The number of events per class and the relative improvement (Rel. Impr %) of the SVGP compared with MLP are also analyzed. The different classes of events present diverse difficulties for classification, depending on the number of instances per class and the variability and specificity of their associated features. EXP and REG are the most challenging types of events. There are two reasons for their lower classification results. First, the number of examples per class is very small compared with the rest of the classes. Second, their spectral and temporal properties are similar to those of other classes, making discrimination more challenging. We can see that this fact is clearly reflected in the f1 score per class in Table I.

SVM-Lin is the worst performing model although it is competitive compared with SVM-SE and RF. Furthermore, as it can be seen, SVGP outperforms every shallow method. In particular, although MLP is better for two classes, SVGP achieves the best accuracy, working especially well on the challenging classes, i.e., EXP and REG. This behavior is a consequence of using nonparametric models against those with a large number of parameters.

The usage of SE-ARD provides a better model convergence, avoiding certain noise introduced by less discriminative or redundant features. Besides, it points out which are these more noisy features and which are the most effective ones. We estimated the length scale for every dimension of the input feature vector. Lower values indicate a higher discriminative power of these features. Fig. 3 shows the length-scale values for the 21 features per event used to feed the classifiers described in Section III-A. For each segment of event (beginning, central, and final part), LPC coefficients 1 and 2 (features 1, 2, 6, 7, 11, and 12 in the feature vector) have the highest relevance. In particular, the shortest length-scale values correspond to the time domain (features 16–18), while LPC coefficients 3–5 in the central segment of signal (features 7–9) present the smallest discriminative relevance. Notice that the most discriminative features have shorter deviation, being relevant across different folds while less discriminative features do not.

Experiments in this section show that, using an SVGP, we are able to outperform widely used shallow methods, such as SVM, RF, or MLP, mainly when the number of events is scarce. Furthermore, information about the discriminative potential of the input features can be extracted when using SVGP and the SE-ARD kernel.

E. Performance of Deep Classifiers

The complexity of seismic events motivates the use of DL although the reduced number of data may make them prone to overfitting. As we will see, the use of deep nonparametric models overcomes this problem. In this section, we compare the best shallow method, i.e., the SVGP, together with its hierarchical extensions, DGPs, i.e., DGP2, DGP3, and DGP4, to the DBN and the sDA reported in [20]. Both DBN and sDA with two and three hidden layers denoted by DBN-H2, DBN-H3, sDA-H2m, and sDA-H3, respectively, are considered. These models use the log loss as the cost function minimizing it with stochastic gradient descent. To avoid overfitting, an early stopping criterion and dropout with $p = 0.20$ are used.

In Table II, we report per class and global f1 scores and accuracy. In addition, we calculated the 95% confidence interval for the global metrics. For the sake of comparison, the relative improvement of the best DGP technique over the best DNN technique is also presented for each class of events. The results confirm the advantages of using deep models for this problem. The reported metrics are better than those in Section III-D except the ones related to the SVGP. SVGP is very competitive for DNNs. It has a lower accuracy (0.9408) than the best DNN, sDA-H2 (0.9432), but the global f1 score

TABLE I

AVERAGED PERFORMANCE IN TEST: F1 SCORE PER CLASS, MACRO-AVERAGE F1 SCORE, AND ACCURACY

	Noise	EXP	REG	COL	VTE	TRE	LPE	*Macro f1*	*Accuracy*
No. Events	1586	278	455	1406	1738	1170	2699		
SVM-Lin	0.9685	0.6639	0.9103	0.9419	0.9342	0.808	0.9275	0.8792±0.0045	0.9155±0.008
SVM-SE	0.9686	0.7186	0.8881	0.9475	0.9339	0.8463	0.9387	0.8917±0.0084	0.9232±0.0076
RF	0.9653	0.7188	0.9013	0.9478	0.941	0.8641	0.9413	0.8971±0.0093	0.928±0.0061
MLP	0.9711	0.7533	0.9003	0.9645	**0.9468**	0.8667	**0.9485**	0.9073±0.0068	0.9373±0.007
SVGP	**0.974**	**0.8002**	**0.9113**	**0.9756**	0.9451	**0.8927**	0.9419	**0.9201**±0.0078	**0.9408** ±0.0027
Rel. Impr. (%)	0.2986	6.2259	1.2218	1.1509	-0.1796	2.9999	-0.6958	1.4108	0.3734

TABLE II

AVERAGED PERFORMANCE IN TEST: F1 SCORE PER CLASS, MACRO-AVERAGE F1 SCORE, AND ACCURACY

	Noise	EXP	REG	COL	VTE	TRE	LPE	*macro F1*	*Accuracy*
No. Events	1586	278	455	1406	1738	1170	2699		
DBN-H2	0.9756	0.7542	0.9143	0.9729	0.9430	0.8898	0.9485	0.9140±0.0065	0.9404±0.0068
sDA-H2	0.9741	0.7778	0.9151	0.9697	0.9484	0.8978	**0.9511**	0.9192±0.006	0.9432±0.0066
DBN-H3	0.97	0.77	0.89	0.97	**0.95**	0.89	0.95	0.91±0.0074	0.9387±0.0069
sDA-H3	0.97	0.78	0.91	0.97	0.94	0.89	0.95	0.92±0.0054	0.9410±0.0068
SVGP	0.974	0.8002	0.9113	0.9756	0.9451	0.8927	0.9419	0.9201±0.0078	0.9408±0.0027
DGP2	0.9789	**0.8391**	0.9175	0.9789	0.9461	**0.9103**	0.9478	**0.9312**±0.0034	0.9477±0.0043
DGP3	0.9765	0.8095	0.9031	0.9723	0.943	0.899	0.9447	0.9211±0.0066	0.9419±0.0058
DGP4	**0.982**	0.8264	**0.9182**	**0.9803**	0.9497	0.9055	0.9472	0.9299±0.0095	**0.9479**±0.0065
Rel. Impr. (%)	0.6560	7.5769	0.3388	0.7606	-0.0316	1.3923	-0.3470	1.2174	0.4983

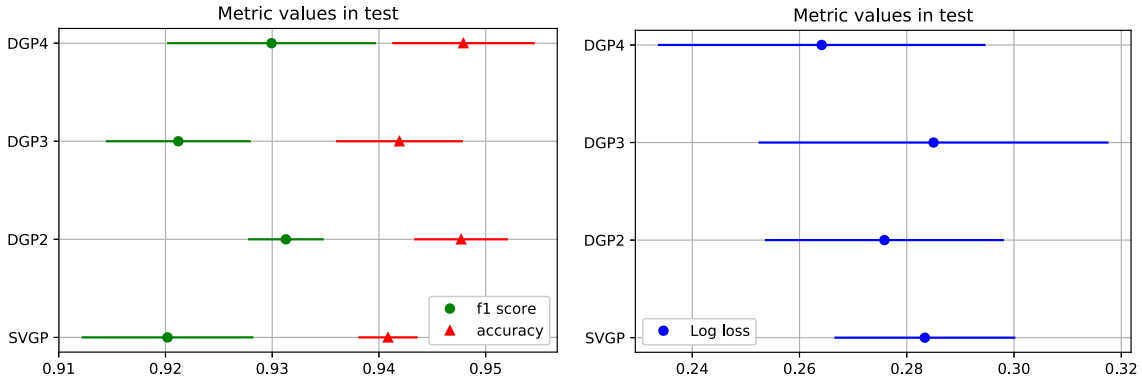
is similar in both, i.e., SVGP (0.9201) and sDA-H3 (0.92). Especially, SVGP outperforms the DNNs for the class EXP showing the capacity of GP models to handle difficult and imbalanced data sets. This fact is confirmed by looking at the best DGP models: DGP2 and DGP4. Both models outperform the rest in accuracy and f1 score obtaining the best global accuracy value and also performing better in difficult and less represented classes. In addition, DGP2 is statistically significant with respect to DNNs since their confidence intervals do not overlap. Regarding the f1 score per class, the best GP-based models, i.e., DGP2 and DGP4, perform remarkably well for difficult classes. Especially, they work notably well in EXP, COL, REG, and TRE, while DNNs only outperform DGPs in the LPE and VTE classes that, together with NOISE, are more easy to identify. It is also worth to point out that DGP2 is the best classifier identifying EXPs (0.8391), with a high relative improvement (7.57%). As it can be observed, the relative improvement is inversely related to the number of events in the class, and in these cases, the difference seems significant. The improvement in the detection of EXP obtained when DGPs are used is very important in monitoring volcanic environments because, together with the LPE and VT, they are often precursors of volcanic activity [2].

In Fig. 4, we depict accuracy, f1 score, and log loss for the GP-based models. Average values of the four cross-validation experiments are depicted with a dot, within an interval line covering the results' standard deviation for the four experiments. This figure provides a better understanding of the results shown in Table II. DGP2 and DGP4 are the best-performing models, while SVGP and DGP3 perform worse. In contrast to DGP2, DGP4 suffers from larger standard deviation values. This higher variance in the results indicates the presence of overfitting in complex models. In this sense, DGP2, with very good performance too, appears to be the model with the greatest generalization capability.

In conclusion, DGPs capture the complex patterns of seismic signals better than DNNs, benefiting from the use of full probabilistic nonparametric models. These results prove the adequacy of the GP-based models for classification of volcano-seismic events. Furthermore, GPs not only perform well globally but also, especially, on these important classes. Finally, DGP2 is the best-performing model with good global accuracy, a reduced variance, and the best result on the most challenging class, i.e., EXPs.

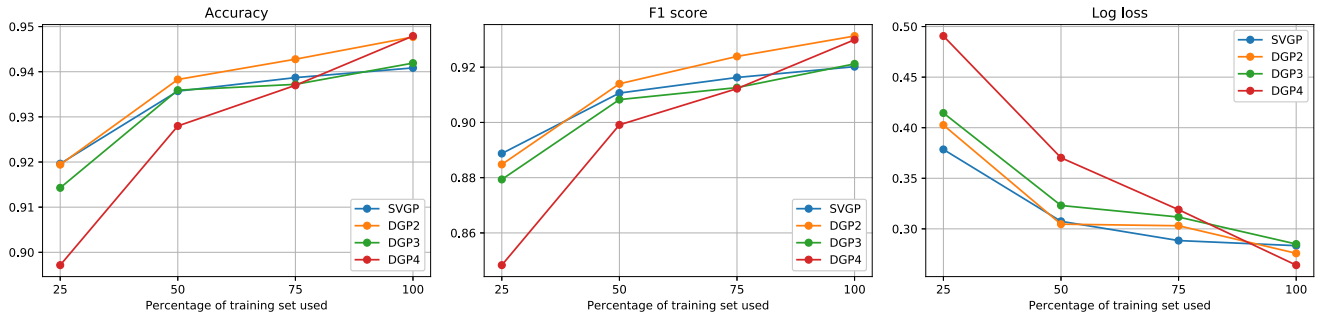
F. Robustness to the Size of the Training Set

In Section III-E, we showed the superiority of GP-based models against DNN ones in test performance. In seismic data, usually, we only have access to a small amount of labeled data, so it is also interesting to analyze the behavior of the studied methods when only a small data set is provided. In this section, we vary the amount of training data available, using 25%, 50%, and 75% of the whole data set. The experiment reveals more about the adequacy of the proposed approach in scenarios where data are scarce. Fig. 5 shows the accuracy, f1 score, and log loss performance of GP-based models. We can clearly see the need of data as the depth of the model increases. When only 25% of the training set is used, DGP4 performs poorly in contrast to the goodness of SVGP and DGP2. The SVGP performance does not improve much with the increase in data; in fact, it is the worst model with the entire data set. In contrast, DGP4 improves enormously as the percentage of data increases. This fact suggests that shallow models are better in scenarios with small data sets, while deeper models, such as DGP4, play an interesting role when more data are available. We also find that DGP2 performs very well through the different experiments achieving very good results both with less and more data. Table III provides an accuracy comparison between the DNN values reported in [20] and those obtained by our GP-based models. Relative improvements of the best



(a) (b)

Fig. 4. Points represent average performance on the test set, and the bars indicate standard deviation. (a) F1 score and accuracy. (b) log loss.



(a) (b) (c)

Fig. 5. Accuracy, f1 score, and log loss metrics varying the percentage of training samples available: 25%, 50%, 75%, and 100% of the training set. (a) Accuracy, (b) f1 score, (c) log loss.

TABLE III

ACCURACY METRIC VARYING THE PERCENTAGE OF TRAINING SAMPLES AVAILABLE AMONG 25%, 50%, AND 75% OF THE TRAINING SET. 100% CORRESPONDS TO THE ENTIRE TRAINING SET

Accuracy	25%	50%	75%	100%
DBN-H2	0.9069	0.9306	0.9283	0.9404
sDA-H2	0.9017	0.9121	0.9283	0.9432
DBN-H3	0.9125	0.922	0.928	0.9387
sDA-H3	0.9077	0.9209	0.9297	0.941
SVGP	**0.9196**	0.9356	0.9386	0.9408
DGP2	0.9194	**0.9382**	**0.9427**	0.9477
DGP3	0.9142	0.9359	0.9372	0.9419
DGP4	0.8971	0.9279	0.9369	**0.9479**
Rel. Impr. (%)	0.7562	0.8167	1.3983	0.4983

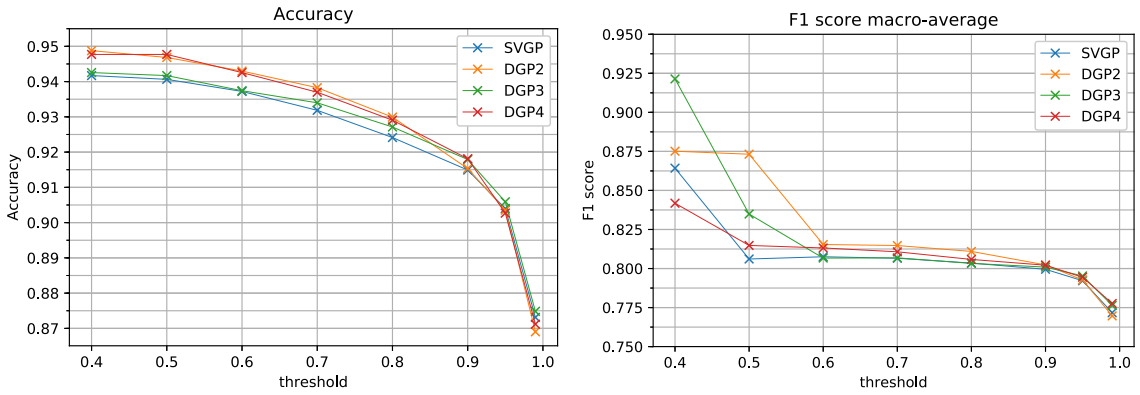
DGP model over the best DNN model are also described. The superiority of GPs is clear. For 25% and 50% of the data, DGP4 has not yet learned a good model, being inferior to the best DNN. Howeverm for 25%, 50%, and 75%, all GP models outperform the best DNN.

As DNNs tend to overfit due to the huge amount of trainable parameters, they are more sensitive to smaller database sizes. In contrast, GPs use prior knowledge that acts as a strong regularization. They learn a good model even when a reduced data set is provided. In summary, as the experiment confirms, GPs perform very well, and better than DNNs, for all data sizes.

G. Evaluating the Confidence in the Predictions

In volcano-seismic applications, it is of paramount importance to analyze the confidence of class predictions. Classification results are often used in early warning tasks: detecting sequences of certain events that are precursors of eruptions; thus, trustable predictions together with a good quantification of their uncertainty are of high interest to design early warning systems. In this section, to analyze the quality of the predictions, we introduce a decision threshold over the probabilities output of the classification systems. By considering as classified only events assigned to a class with probability greater than the threshold and varying this threshold, we can increase the confidence of the system. Two studies are performed.

1) First, we study accuracy and f1 score when different threshold values are used (0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0,95 y 0.99). The results are shown in Fig. 6. As we increase the threshold, we are predicting fewer examples correctly; therefore, accuracy and f1 score decrease. We can see that most samples are predicted with at least a 0.99 probability, and we misclassify only 7% if we change the threshold from 0.4 to 0.99. Note that 0.99 is a very demanding threshold being most samples predicted with very high probability (close to 1). The faster decrease of the f1 score suggests that there is more uncertainty in the less represented classes.



(a) (b)

Fig. 6. Accuracy and f1 score metrics varying the classification probability threshold. A sample is predicted if the output probability of the highest class probability is higher than the selected threshold; otherwise, this sample is unclassified. (a) Accuracy and (b) f1 score macro-average.

In accuracy, for low thresholds, i.e., 0.4, 0.5, and 0.6, DGP2 and DGP4 classify more events correctly than SVGP and DGP3, but, with higher thresholds, this difference decreases (or is reduced). Regarding the f1 score, we observe the same behavior. Thus, the number of events classified with high probabilities is almost the same in all cases, but DGP2 and DGP4 are able to give more confidence to doubtful samples.

2) For a complete understanding of the classifier confidence, Fig. 7 shows the distribution of the predicted probabilities per class. The X-axis represents the probability of belonging to the class predicted. The Y-Axis represents the cumulative density function of these probabilities for each class of events. Comparing the probability CDFs for different models, the figure provides information about how trustable the different classifications are.

   First, we confirm that few samples are predicted with very low probability; indeed, most of the predictions are close to 1. This fact confirms that the models are confident in the predictions. They define good decision boundaries and identify every class well. All classifiers have a similar performance except for EXP and TRE; as we saw in Section III-E, both are the most difficult types of events. For these events, DGP2 and DGP4 perform better than SVGP and DGP3. This fact matches the log loss reported in Fig. 4.

   Titos [20, Fig. 4] reported the same cumulative density functions of classification probabilities values for DNNs. We can see that GP-based methods outperform DNN ones in this experiment. For example, in EXP, the difference is quite clear; 50% is predicted with a 0.9 probability or more by the DNNs, and in contrast, GPs predicted more than 60% with high probability, i.e., 0.9 or more.

In conclusion, in this work, we observed that, for this database, the probabilities given by the GPs are more trustworthy than the ones provided by the DNNs, and the best-performing GP-based model, i.e., DGP2, is also the most confident.

IV. Conclusion

In this work, we have introduced to the seismic community the usage of SVGPs, their hierarchical extension, and DGPs for automatic volcano-seismic event classification. We tested them on the seismic database recorded at Volcán de Fuego de Colima.

Due to the complexity of this problem, state-of-the-art methods are based on hierarchical deep models, i.e., DNNs. However, they require more data than usually available. The obtained results indicate that SVGPs outperform all the shallow classifiers. Moreover, they are competitive with DNNs. The two-layer DGP outperforms DNNs avoiding overfitting. It attains both good accuracy and f1 score and performs better than DNNs on difficult classes.

We have proven the adequacy of GPs with additional experiments. The experiments indicate that they can still learn good models even when the database is small. When data are scarce, SVGP was the best-performing method. Besides, with more data, deeper models, such as four-layer DGPs, are an interesting option with promising results. In general, the two-layer DGP performed very well through different percentages of training data. Finally, we carried out an exhaustive study on the prediction confidence. GP-based methods obtained probabilities closer to 1 than DNNs.

These experiments suggest that GP-based methods are able to classify very well seismic events, especially interesting classes, such as EXP, REG, and LPE, even when data are scarce. Besides, they take into account the model uncertainty, being a trustworthy system for volcanologists. In short, we have shown that GPs and DGPs can be applied with success to seismic problems.

Appendix
Detailed Introduction of Gaussian Processes and Deep Gaussian Processes

In this appendix, we provide a more detailed introduction to the use of GPs and DGPs for multiclass classification problems. We explain their probabilistic formulation, provide some intuition and examples of them, and describe how

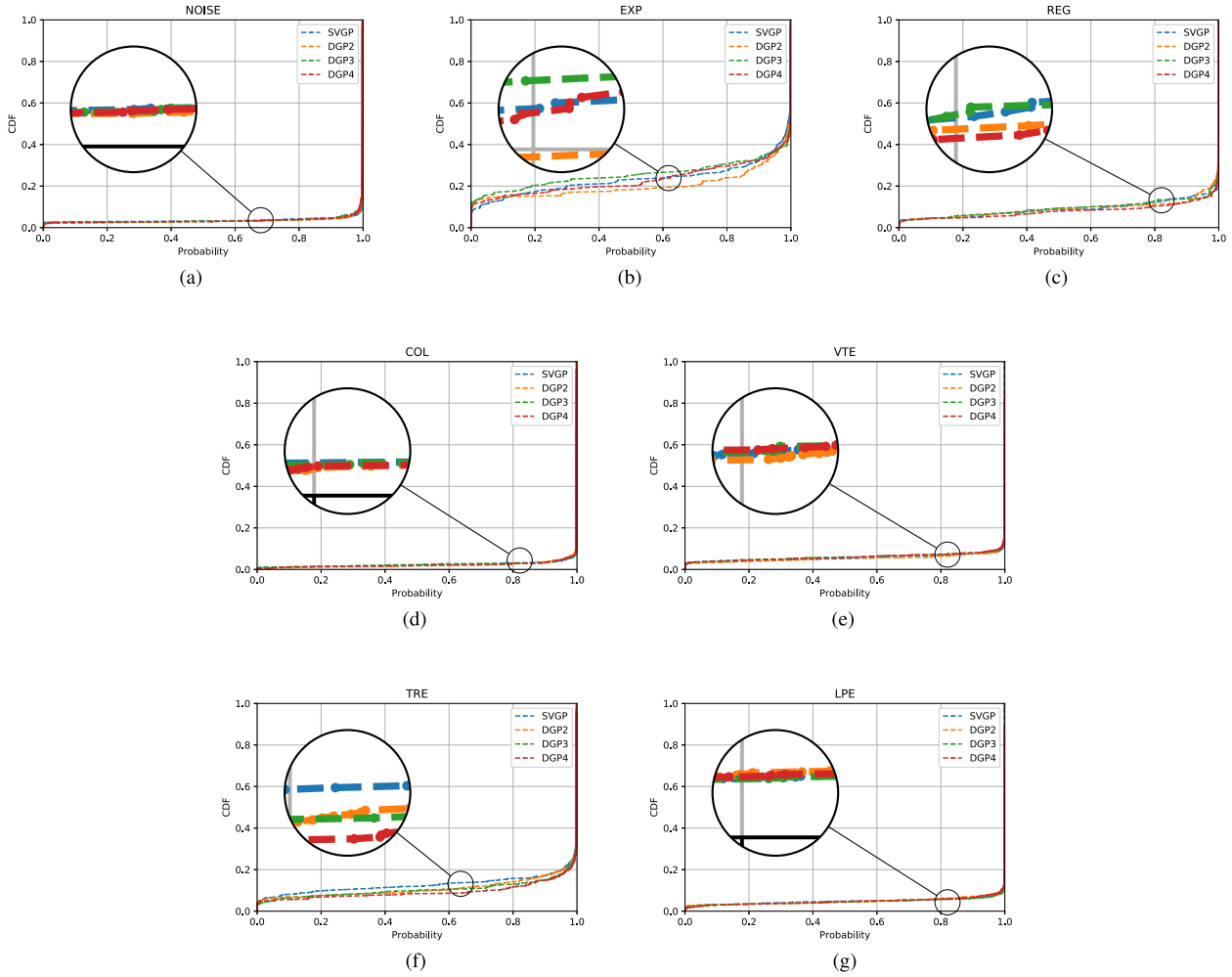


Fig. 7. Cumulative distribution function of the probabilities given by the GP classifiers per class. In the Y-axis, we represent the proportion of samples of that class with a certain predicted probability or less. In the X-axis, we represent these predicted probabilities. (a) NOISE. (b) EXP. (c) REG. (d) COL. (e) VTE. (f) TRE. (g) LPE.

inference is carried out. An in-depth study of the inference methods followed here can be found in [34] for GPs and in [30] for DGPs.

A multiclass classification problem with K classes consists of N labeled instances $\{(\mathbf{x}_n, y_n)\}_{n=1}^N$, where $\mathbf{x}_n \in \mathbb{R}^D$ is the feature vector and $y_n \in \{1, \ldots, K\}$ is the class label of the nth instance. We define the $N \times D$ matrix $\mathbf{X}$ as the feature matrix where, in the nth row, we have the feature vector of the nth instance. In this work, the features ($D = 21$) are extracted from the raw signal, and more information about them is provided in Section III-A. We also define $\mathbf{y}$ the vector that gathers the labels of the samples. Once the supervised classifier is trained, it is able to provide the class label y_* for any unseen instance $\mathbf{x}_*$.

A. Single-Layer GPs

For each instance $\mathbf{x}_n$, its label y_n is modeled using K latent variables $\mathbf{f}_{n,:} = \{f_k(\mathbf{x}_n)\}_{k=1}^K$ through a specific likelihood $\mathrm{p}(\mathbf{y}|\mathbf{f}_{n,:})$. The likelihood squashes the values of the latent variable defined in $\mathbb{R}$ to the $[0, 1]$ interval. Notice that this likelihood plays a similar role as the output neurons play in DNNs. For example, the so extended softmax function can be used here. In this work, we utilize the robust max likelihood, which prevents overfitting in GPs. It is defined by

$$\mathrm{p}(y_n = k|\mathbf{f}_{n,:}) = \begin{cases} 1-\varepsilon, & k = \underset{1\le j\le K}{\arg\max}\ \mathbf{f}_{n,j} \\ \dfrac{\varepsilon}{K-1}, & \text{otherwise} \end{cases} \tag{5}$$

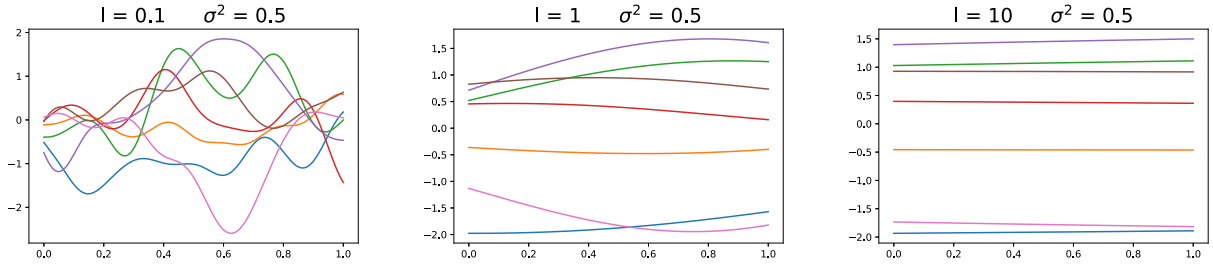
with $k \in \{1, \ldots, K\}$ and $1-(1/K) > \varepsilon > 0$, which is usually fixed to a small value; in this work, it was fixed to 10^{-3}. For simplicity, we denote the latent variables by $f_k(\mathbf{x}_n) = f_{n,k}$.

We factorize the likelihood assuming that the class labels are independent for the different samples

$$\mathrm{p}(\mathbf{y}|\mathbf{F}) = \prod_{n=1}^{N} \mathrm{p}(y_n|\mathbf{f}_{n,:}) \tag{6}$$

where $\mathrm{p}(y_n|\mathbf{f}_{n,:})$ is given by (5). The $N \times K$ matrix $\mathbf{F}$ gathers the K latent variables for the N instances. The (n, k) term corresponds to the kth latent variable for the nth instance. The nth row of $\mathbf{F}$ is denoted by $\mathbf{f}_{n,:}$ and the kth column by $\mathbf{f}_k$.

Having defined the observation model, we now turn our attention to the definition of the prior model on $\mathbf{F}$. Notice that,



(a) (b) (c)

Fig. 8. 1-D example of a GP. We draw several samples from a GP with an SE kernel varying the length scale. Shorter values of the length scale l produce wriggly curves, while larger values produce flat functions. (a) $l = 0.1$, (b) $l = 1$, (c) $l = 10$.

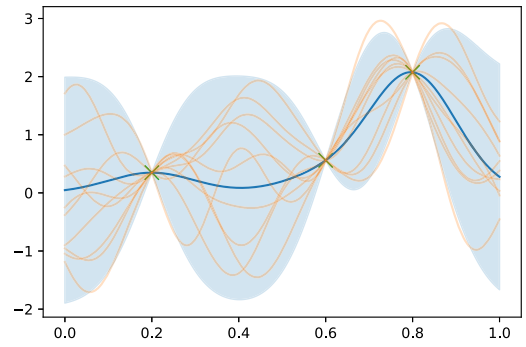


Fig. 9. 1-D example of a GP with an SE kernel ($l = 0.1$ and $\sigma^2 = 1$). We have observed the values in $x_1 = 0.2$, $x_2 = 0.6$, and $x_3 = 0.8$. Then, we predict in 100 unobserved points X_* of the [0, 1] interval given these observations. We draw $\mathrm{p}(F_*|X_*, X, F, \boldsymbol{\Theta})$ with $X = \{x_1, x_2, x_3\}$ and $F = \{f(x_1), f(x_2), f(x_3)\}$: the blue line is the mean, and the blue shadow is the 0.95 confidence interval. We also draw several samples from this distribution in orange. Observe that almost all samples are contained in the confidence interval.

at observation level, if the class of the nth sample is k, $f_{n,k}(\mathbf{x}_n)$ is larger than $f_{n,j}(\mathbf{x}_n)$, $j \neq k$ and that we are assuming that *a priori* $\mathbf{f}_k$ and $\mathbf{f}_j$, $j \neq k$, are independent. Thus, we need to model now the *a priori* behavior of each $\mathbf{f}_i, i = 1, \ldots, K$. We use a GP to define an *a priori* independent distribution for each column component of the latent matrix. A GP is an infinite collection of random variables in which every finite subset is Gaussian distributed. It can be seen as a prior over functions. Thus, we assume that the columns of the latent variable $\mathbf{F}$, $\{\mathbf{f}_k\}_{k=1}^K$, follow independent GP priors. For every k, it imposes that $\{f_{n,k}\}_{n=1}^N$ follow jointly a Gaussian distribution $\mathcal{N}(\mathbf{f}_k|\mathbf{0}, \mathbf{K_{XX}})$, where the covariance matrix is obtained using a kernel function $k(\cdot, \cdot)$ [24]. We can write the prior distribution of the latent function as

$$\mathrm{p}(\mathbf{F}|\boldsymbol{\Theta}, \mathbf{X}) = \prod_{k=1}^{K} \mathrm{p}(\mathbf{f}_k|\boldsymbol{\Theta}, \mathbf{X}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{f}_k|\mathbf{0}, \mathbf{K_{XX}}) \tag{7}$$

where $\boldsymbol{\Theta}$ are the kernel hyperparameters. The covariance matrix $\mathbf{K_{XX}} = \mathbf{K}(\mathbf{X}, \mathbf{X}) = (k(\mathbf{x}_i, \mathbf{x}_j))_{i,j}$ encodes the properties of the desirable function (e.g., smoothness).

In this work, we use the SE kernel

$$k_{\mathrm{SE}}(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 \exp\left(\frac{-||\mathbf{x}_i - \mathbf{x}_j||^2}{2l^2}\right). \tag{8}$$

This kernel has a great power of representation, and it is used in many different scenarios [24]. In this case, we have to estimate the length scale l and variance σ^2 hyperparameters. Note that functions drawn from a GP with an SE kernel are infinitely differentiable leading to smooth functions that are desirable in most problems. In Fig. 8, assuming that $x_1, \ldots, x_N$ are 100 points evenly distributed in the interval [0, 1], we show several samples of a GP with different elections of the length scale, and we can notice that it controls the level of smoothness. Larger values of this parameter produce flat functions, while shorter values lead to wriggly functions. Assuming that the GP has been observed only at $x_1 = 0.2$, $x_2 = 0.6$, and $x_3 = 0.8$, we show, in Fig. 9, the observed values together with the predicted values $f(X_*)$ for X_* being 100 points evenly distributed in the [0, 1] interval. The use of a GP imposes that $f(X_*), f(x_1,), \ldots, f(x_3)$ are jointly Gaussian from which we can obtain the distribution of $f(X_*)$ given $f(x_1,), \ldots, f(x_3)$. We also include their 0.95 confidence intervals.

The joint distribution of the probabilistic framework defined here is given by

$$\mathrm{p}(\mathbf{y}, \mathbf{F}, \mathbf{X}|\boldsymbol{\Theta}) = \underbrace{\mathrm{p}(\mathbf{y}|\mathbf{F})}_{\text{likelihood}} \underbrace{\mathrm{p}(\mathbf{F}|\mathbf{X}, \boldsymbol{\Theta})}_{\text{GP prior}}. \tag{9}$$

Approximate inference methods, such as the Laplace method or expectation propagation, have a computational cost of $\mathcal{O}(KN^3)$ because they involve the inversion of an $N \times N$ dimensional matrix. To amend this problem, we use the sparse approximation of GPs [34]. We define $M \ll N$ inducing points for each GP. These inducing points are latent variables, and they are the values of the GP realization at the inducing point locations $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_M\} \subset \mathbb{R}^D$. We gather them in the $M \times K$ matrix $\mathbf{U}$. The (m, k) term corresponds to the kth latent variable of the mth inducing point. The mth row of $\mathbf{U}$ is denoted by $\mathbf{u}_{m,:}$ and the kth column by $\mathbf{u}_k$. As we have indicated, these inducing points can be seen as $\mathbf{U} = \mathbf{F}(\mathbf{Z})$. We are summarizing the value of the true latent function through the inducing points, so it is important to optimize on their location. It is expected that these optimal locations will end up close to informative places as the decision boundaries. The probabilistic model of the sparse approach is given by

$$\mathrm{p}(\mathbf{y}, \mathbf{F}, \mathbf{U}|\boldsymbol{\Theta}) = \underbrace{\mathrm{p}(\mathbf{y}|\mathbf{F})}_{\text{likelihood}} \underbrace{\mathrm{p}(\mathbf{F}|\mathbf{U}, \boldsymbol{\Theta})\mathrm{p}(\mathbf{U}|\boldsymbol{\Theta})}_{\text{GP prior}}. \tag{10}$$

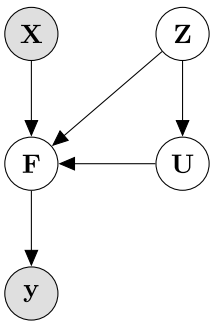


Fig. 10. Probabilistic graphical model of an SVGP. Dark circles stand for observed variables, while light circles stand for latent variables.

Notice that

$$\mathrm{p}(\mathbf{y}, \mathbf{F}|\boldsymbol{\Theta}) = \int \mathrm{p}(\mathbf{y}|\mathbf{F})\mathrm{p}(\mathbf{F}|\mathbf{U}, \boldsymbol{\Theta})\mathrm{p}(\mathbf{U}|\boldsymbol{\Theta})\mathrm{d}\mathbf{U} \tag{11}$$

and so the abovementioned factorization does not modify the modeling. Fortunately, it provides us with a tool to perform tractable inference. We show the probabilistic graphical model using inducing points in Fig. 10.

In this work, we follow the SVGP [34]. It will allow to estimate the model parameters $\boldsymbol{\Theta}$ and approximate the posterior distribution $\mathrm{p}(\mathbf{F}, \mathbf{U}|\mathbf{y}, \boldsymbol{\Theta})$ by the distribution $\mathrm{q}(\mathbf{F}, \mathbf{U})$. Using the joint distribution and Jensen's inequality, we obtain the well-known evidence lower bound (ELBO)

$$\log \mathrm{p}(\mathbf{y}|\boldsymbol{\Theta}) \geq \int \mathrm{q}(\mathbf{F}, \mathbf{U}) \log \frac{\mathrm{p}(\mathbf{y}, \mathbf{F}, \mathbf{U}|\boldsymbol{\Theta})}{\mathrm{q}(\mathbf{F}, \mathbf{U})}\mathrm{d}\mathbf{U}\mathrm{d}\mathbf{F}. \tag{12}$$

Notice that this bound is valid for every $\mathrm{q}(\mathbf{F}, \mathbf{U})$ distribution. It is straightforward to see that maximizing the ELBO is equivalent to minimize the Kullback–Leibler divergence between $\mathrm{q}(\mathbf{F}, \mathbf{U})$ and $\mathrm{p}(\mathbf{F}, \mathbf{U}|\mathbf{y}, \boldsymbol{\Theta})$. The SVGP approximation utilizes the following parametric form for q:

$$\mathrm{q}(\mathbf{F}, \mathbf{U}) = \mathrm{q}(\mathbf{F}|\mathbf{U}, \Theta)\mathrm{q}(\mathbf{U}) \tag{13}$$

$$\mathrm{q}(\mathbf{F}|\mathbf{U}, \boldsymbol{\Theta}) = \mathrm{p}(\mathbf{F}|\mathbf{U}, \boldsymbol{\Theta}) \tag{14}$$

$$\mathrm{q}(\mathbf{U}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{u}_k|_k, \mathbf{S}_k). \tag{15}$$

Then, the ELBO can be rewritten as

$$\begin{aligned}
&\log \mathrm{p}(\mathbf{y}|\boldsymbol{\Theta}) \\
&\geq \int \mathrm{q}(\mathbf{U})\mathrm{p}(\mathbf{F}|\mathbf{U}) \log \frac{\mathrm{p}(\mathbf{y}|\mathbf{F})\mathrm{p}(\mathbf{F}|\mathbf{U})\mathrm{p}(\mathbf{U})}{\mathrm{p}(\mathbf{F}|\mathbf{U})\mathrm{q}(\mathbf{U})}\mathrm{d}\mathbf{U}\mathrm{d}\mathbf{F} \\
&= \mathbb{E}_{\mathrm{p}(\mathbf{F}|\mathbf{U})\mathrm{q}(\mathbf{U})} \log \mathrm{p}(\mathbf{Y}|\mathbf{F}) + \mathbb{E}_{\mathrm{q}(\mathbf{U})}\left(\frac{\mathrm{p}(\mathbf{U})}{\mathrm{q}(\mathbf{U})}\right) \\
&= \sum_{n=1}^{N} \mathbb{E}_{\mathrm{q}(\mathbf{f}_{n,:})} \log \mathrm{p}(\mathbf{y}_i|\mathbf{f}_{n,:}) - \sum_{k=1}^{K} \mathrm{KL}(\mathrm{q}(\mathbf{u}_k)||\mathrm{p}(\mathbf{u}_k))
\end{aligned} \tag{16}$$

where KL is the Kullback–Leibler divergence. The derivation in (16) allows to see the ELBO as the sum of two terms: the first one is a fidelity term imposing that the latent classifier must classify well, and the second one is a regularization term over the latent variable in the inducing points. Our final goal then becomes to find the optimal kernel hyperparameters $\tilde{\boldsymbol{\Theta}}$, inducing locations $\tilde{\mathbf{Z}}$ and variational parameters of $\mathrm{q}(\mathbf{U})$, i.e., $\tilde{\mathbf{m}}_k, \tilde{\mathbf{S}}_k$, by maximizing the ELBO in (16). Furthermore, since the ELBO factorizes over the instances, we can use minibatches for optimizing this function reducing the computational cost; in this case, considering that $M < N_b$, the computational cost is $\mathcal{O}(N_b M^2 K)$, where N_b is the minibatch size.

Once the ELBO is optimized and the variational parameters computed, we can make predictions on an unseen test sample $\mathbf{x}_*$. The value of the latent variable $\mathbf{f}_*$ on this point $\mathbf{x}_*$ is given by

$$\begin{aligned}
\mathrm{p}(f_{*,k}|\mathbf{x}_*, \tilde{\boldsymbol{\Theta}}, \mathbf{X}, \mathbf{y}) &= \int \mathrm{p}(f_{*,k}|\mathbf{u}_k)\mathrm{p}(\mathbf{u}_k|\tilde{\boldsymbol{\Theta}})d\mathbf{u}_k \\
&\approx \mathbb{E}_{\mathrm{q}(\mathbf{u}_k)}\mathrm{p}(f_{*,k}|\mathbf{u}_k) \\
&= \mathcal{N}(f_{*,k}|\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})
\end{aligned} \tag{17}$$

where the mean and the covariance matrix are defined by

$$\tilde{\boldsymbol{\mu}} = \mathbf{K}_{\mathbf{x}_*\tilde{\mathbf{Z}}}\mathbf{K}_{\tilde{\mathbf{Z}},\tilde{\mathbf{Z}}}^{-1}\tilde{k} \tag{18}$$

$$\tilde{\boldsymbol{\Sigma}} = k_{\mathbf{x}_*\mathbf{x}_*} + \mathbf{K}_{\mathbf{x}_*\tilde{\mathbf{Z}}}\mathbf{K}_{\tilde{\mathbf{Z}},\tilde{\mathbf{Z}}}^{-1}(\tilde{\mathbf{S}}_k - \mathbf{K}_{\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}})\mathbf{K}_{\tilde{\mathbf{Z}},\tilde{\mathbf{Z}}}^{-1}\mathbf{K}_{\tilde{\mathbf{Z}}\mathbf{x}_*}\big). \tag{19}$$

Finally, the class label is obtained using

$$\mathrm{p}(\mathbf{y}_*) = \int \mathrm{p}(\mathbf{y}_*|\mathbf{f}_*)\mathrm{p}(\mathbf{f}_*|\mathbf{x}_*, \tilde{\boldsymbol{\Theta}}, \mathbf{X}, \mathbf{y})d\mathbf{f}_*. \tag{20}$$

This integral is intractable and it can be computed using numerical algorithms, e.g., the Gaussian–Hermite quadrature.

In Fig. 11, we depict a 1-D binary toy example to provide a better understanding of the SVGP model. The observation model is

$$\mathrm{p}(y_n|f(x_n)) = \left(\frac{1}{1+e^{-f(x_n)}}\right)^{y_n}\left(1 - \frac{1}{1+e^{-f(x_n)}}\right)^{1-y_n} \tag{21}$$

with $y_n \in \{0, 1\}, x_n \in \mathbb{R}$. Notice that, here, we only have one SVGP. The blue dots are class 0 and 1 observations, and they have been observed at $x \in [0, 1]$. In Fig. 11(a), the blue line represents the mean of the posterior latent function distribution, and the blue shadow represents the confidence interval. The wider this shadow, the more the uncertainty. We can see that there is more uncertainty in the middle of the interval because there are no observations there. This latent function takes values in $\mathbb{R}$, so it has to be squashed into the $[0, 1]$ interval using the likelihood. In Fig. 11(b), the black line goes from 0 to 1 and corresponds to the value of $\mathrm{p}(\mathbf{y}_*)$ for $y_* = 1$ in (20). Notice how this value takes into account all the possible values of f_*.

B. Deep Gaussian Processes

In this section, we detail the hierarchical extension of SVGP. Roughly speaking, the idea behind DGPs is to stack several SVGPs. If we use the output of one SVGP as the input of another SVGP, and we repeat this procedure L times that we define the $(L + 1)$ layer. DGPs were first introduced in [29].

As it happens to SVGP, exact inference is also intractable for DGPs. In this work, we follow the doubly stochastic inference proposed in [30]. We introduce, at each layer l, M inducing points $\mathbf{U}^l$ at inducing locations $\mathbf{Z}^{l-1}$. The joint distribution of

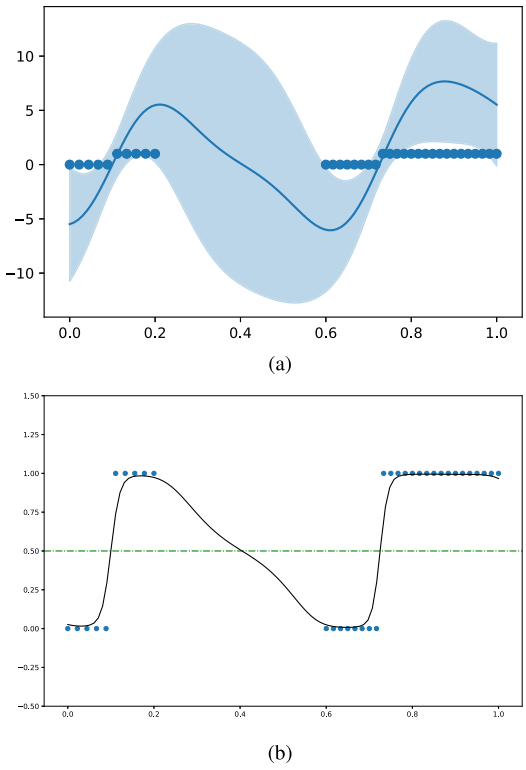


Fig. 11. 1-D binary classification problem. The blue points represent the observations. In (a), we draw $\mathrm{p}(f_*)$: the blue line is the mean, and the blue shadow is the 0.95 confidence interval on the predictions. The classifier has more uncertainty in the region where there are no observations. In (b), we squash the latent function to the [0, 1] interval, and the black line is $\mathrm{p}(y_* = 1)$.

the probabilistic framework defined here is given by

$$\mathrm{p}(\mathbf{y}, \{\mathbf{F}^l, \mathbf{U}^l\}_{l=1}^L) = \underbrace{\prod_{n=1}^N \mathrm{p}(y_n|f_n^L)}_{\text{likelihood}} \times \underbrace{\prod_{l=1}^L \mathrm{p}(\mathbf{F}^l|\mathbf{U}^l;\mathbf{F}^{l-1},\mathbf{Z}^{l-1})\mathrm{p}(\mathbf{U}^l;\mathbf{Z}^{l-1})}_{\text{DGP prior}} \quad (22)$$

We consider $\mathbf{F}^0 = \mathbf{X}$, and each factor in the product is the joint distribution over $(\mathbf{F}^l, \mathbf{U}^l)$ of an SVGP in the inputs $(\mathbf{F}^{l-1}, \mathbf{Z}^{l-1})$ but rewritten with the conditional probability given $\mathbf{U}^l$. We introduce here the semicolon notation to clarify which are the inputs in the equations. We also consider the same amount of inducing points in every layer, but notice that the hidden size of each layer can be different. $\mathbf{F}^l$ and $\mathbf{U}^l$ are $N \times D^l$ and $M \times D^l$ matrices, respectively. In this case, $\mathbf{Z}^{l-1}$ is a $M \times D^{l-1}$ matrix. We show the graphical probabilistic model in Fig. 12 that illustrates the hierarchical construction of this architecture.

Following the same approach used in the single-layer case, we use variational inference to find a posterior distribution approximation $\mathrm{q}(\{\mathbf{F}^l, \mathbf{U}^l\}_{l=1}^L)$

$$\mathrm{q}\big(\{\mathbf{F}^l, \mathbf{U}^l\}_{l=1}^L\big) = \prod_{l=1}^L \mathrm{p}(\mathbf{F}^l|\mathbf{U}^l;\mathbf{F}^{l-1},\mathbf{Z}^{l-1})\mathrm{q}(\mathbf{U}^l) \quad (23)$$

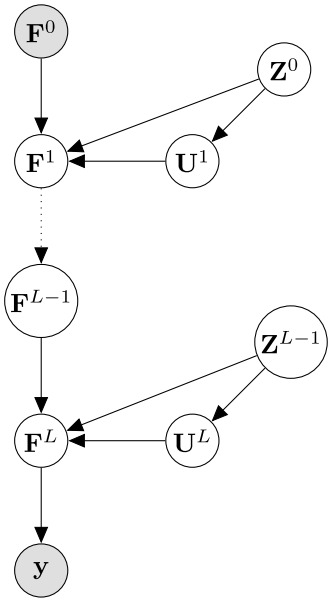


Fig. 12. Probabilistic graphical model of a DGP with L layers. Dark circles stand for observed variables, while light circles stand for latent variables. The dotted arrow refers to the inductive process for building the general deep model.

where we impose the factorization $\mathrm{q}(\mathbf{U}^l) = \mathcal{N}(\mathbf{U}^l|^l, \mathbf{S}^l)$. The ELBO can then be written as

$$\begin{aligned}
&\log \mathrm{p}(\mathbf{y}) \\
&\geq \int \prod_{l=1}^L \mathrm{p}(\mathbf{F}^l|\mathbf{U}^l;\mathbf{F}^{l-1},\mathbf{Z}^{l-1})\mathrm{q}(\mathbf{U}^l) \\
&\quad \times \log \frac{\prod_{n=1}^N \mathrm{p}(y_n|\mathbf{f}_{n,:}^L)\prod_{l=1}^L \mathrm{p}(\mathbf{F}^l|\mathbf{U}^l;\mathbf{F}^{l-1},\mathbf{Z}^{l-1})\mathrm{p}(\mathbf{U}^l;\mathbf{Z}^{l-1})}{\prod_{l=1}^L \mathrm{p}(\mathbf{F}^l|\mathbf{U}^l;\mathbf{F}^{l-1},\mathbf{Z}^{l-1})\mathrm{q}(\mathbf{U}^l)} \\
&\quad \times \prod_{l=1}^L \mathrm{d}\mathbf{U}^l \mathrm{d}\mathbf{F}^l \\
&= \sum_{n=1}^N \mathbb{E}_{\mathrm{q}(\mathbf{f}_{n,:}^L)}[\log \mathrm{p}(y_n|\mathbf{f}_{n,:}^L)] - \sum_{l=1}^L \mathrm{KL}(\mathrm{q}(\mathbf{U}^l)||\mathrm{p}(\mathbf{U}^l;\mathbf{Z}^{l-1})).
\end{aligned} \quad (24)$$

Now, we also estimate the model parameters for every layer, the variational parameters of $\mathrm{q}(\mathbf{U}^l)$, and the inducing point locations $\mathbf{Z}^{l-1}$. Again, the first term corresponds to a fidelity term and the second one to a regularization of the latent variable at each layer. In this case, the second term is tractable since it is the KL divergence between Gaussians. However, the first term involves the marginals of the posterior at the last layer, $\mathrm{q}(\mathbf{f}_{n,:}^L)$, which is analytically intractable. Fortunately, it can be sampled efficiently using univariate Gaussians.

Marginalizing out the inducing points in (23), the posterior distribution for the GP layers $\{\mathbf{F}^l\}_{l=1}^L$ becomes

$$\mathrm{q}(\{\mathbf{F}^l\}_{l=1}^L) = \prod_{l=1}^L \mathrm{q}(\mathbf{F}^l|^l, \mathbf{S}^l;\mathbf{F}^{l-1},\mathbf{Z}^{l-1}) = \prod_{l=1}^L \mathcal{N}(\mathbf{F}^l|\tilde{\boldsymbol{\mu}}^l, \tilde{\boldsymbol{\Sigma}}^l) \quad (25)$$

where $[\tilde{\boldsymbol{\mu}}^l]_n = \mu_{l,\mathbf{Z}^{l-1}}(\mathbf{f}_{n,:}^{l-1})$ and $[\tilde{\boldsymbol{\Sigma}}^l]_{ij} = \Sigma_{\mathbf{S}^l,\mathbf{z}^{l-1}}(\mathbf{f}_{i,:}^{l-1}, \mathbf{f}_{j,:}^{l-1})$. The specific form of the functions $\mu_{l,\mathbf{Z}^{l-1}}$ and $\Sigma_{\mathbf{S}^l,\mathbf{z}^{l-1}}$ can be

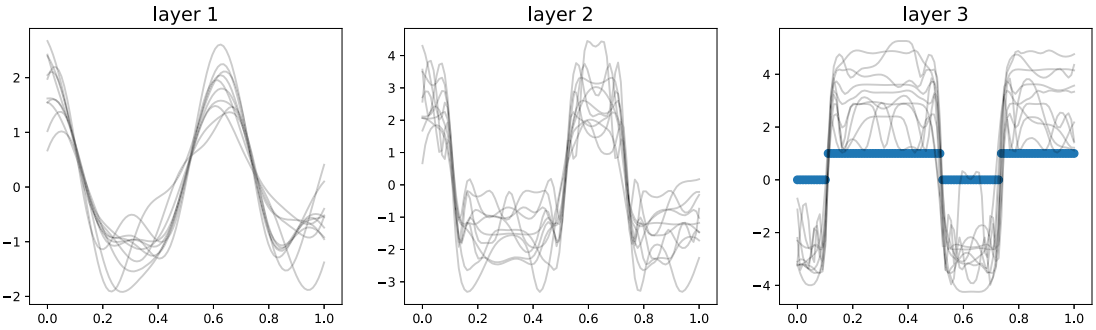


Fig. 13. Samples from the posterior distribution of the latent function at every layer of a three-layer DGP on a 1-D binary classification problem. Every layer is endowed with an SE kernel. The observations are described by the blue points on the third picture. Every layer provides a higher level of abstraction producing more complex patterns.

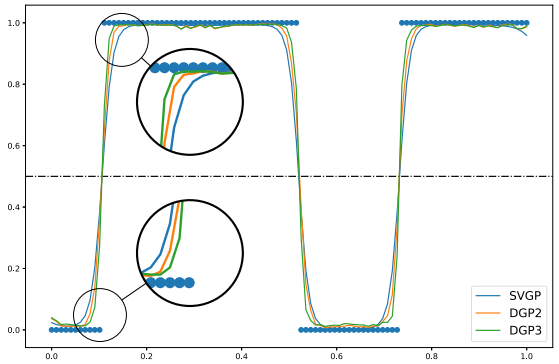


Fig. 14. Comparison of an SVGP, a two-layer DGP (DGP2), and a three-layer DGP (DGP3) on a 1-D binary classification problem. Deeper models are able to capture better the decision boundary (see the zoomed-in areas).

found in [30, eqs. (7) and (8)]. Notice that we are able to compute the nth marginal at each layer $\mathcal{N}(\mathbf{f}_{n,:}^l|[\tilde{\boldsymbol{\mu}}^l]_n, [\tilde{\boldsymbol{\Sigma}}^l]_{nn})$ since it only depends on the corresponding nth input of the previous layer. Thus, taking a sample of $\mathrm{q}(\mathbf{f}_{n,:}^L)$ is straightforward, and we have to recursively sample from the first to the last layer $\hat{\mathbf{f}}_{n,:}^1 \to \hat{\mathbf{f}}_{n,:}^2 \to \cdots \to \hat{\mathbf{f}}_{n,:}^L$. Especially, we first sample from $\boldsymbol{\varepsilon}_n^l \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{D^l})$, and then, for $l = 1, \ldots, L$, we sample

$$\hat{\mathbf{f}}_{n,:}^l = \mu_{l,\mathbf{Z}^{l-1}}\left(\hat{\mathbf{f}}_{n,:}^{l-1}\right) + \boldsymbol{\varepsilon}_n^l \cdot \sqrt{\Sigma_{\mathbf{S}^l,\mathbf{Z}^{l-1}}\left(\hat{\mathbf{f}}_{n,:}^{l-1}, \hat{\mathbf{f}}_{n,:}^{l-1}\right)}. \tag{26}$$

In summary, the expectation $\mathbb{E}_{\mathrm{q}(\mathbf{f}_{n,:}^L)}[\log \mathrm{p}(y_n|\mathbf{f}_{n,:}^L)]$ in the ELBO [see (24)] can be approximated with a Monte Carlo sample generated using (26). Since the ELBO factorizes across data points and the samples can be drawn independently for each point n, scalability is achieved through subsampling the data in minibatches. The complexity to evaluate the ELBO and its gradients is $\mathcal{O}(N_b M^2 \sum_{l=1}^L D^l)$. Notice how the number of layers, and especially the hidden dimension of each one, increases the computational cost in comparison to a single layer SVGP.

Once the ELBO is optimized, we can make predictions on an unseen test sample $\mathbf{x}_*$. The value of the latent variable $\mathbf{f}_{*,:}^L$ can be approximated by taking S samples[1] from the posterior up to the $(L-1)$th layer using $\mathbf{x}_*$ as the initial input. This yields a set $\{\mathbf{f}_{*,:}^{L-1}(s)\}_{s=1}^S$. Then, the density over $\mathbf{f}_{*,:}^L$ is given by the Gaussian mixture (recall that all the terms in (25) are Gaussians)

$$\mathrm{q}(\mathbf{f}_{*,:}^L) \approx \frac{1}{S}\sum_{s=1}^{S} \mathrm{q}(\mathbf{f}_{*,:}^L|^L, \mathbf{S}^L; \mathbf{f}_{*,:}^{L-1}(s), \mathbf{Z}^{L-1}). \tag{27}$$

The code to perform DGP inference and prediction is integrated within GPflow (a GP framework built on top of Tensorflow) and is publicly available.[2]

To illustrate the intuition behind DGPs, we show, in Fig. 13, samples from a three-layer DGP on a 1-D binary classification problem. We equipped each layer with an SE kernel and drew samples from the posterior distribution of the latent function at every layer. The SE kernel produces very smooth functions in the first layer. However, the concatenation of these simple functions produces more complex functions as we increase the depth. In the last layer, it captures very sophisticated patterns combining flat regions with high-variability ones. These patterns cannot be captured by a shallow GP with a stationary kernel. A comparison between a shallow SVGP and DGPs in this problem is shown in Fig. 14. Both models perform very well because the problem is very simple; however, we can still notice one of the main differences between SVGPs and DGPs. Deeper models are able to make an abrupter jump defining better the decision boundary (see the zoomed-in areas). In this case, the SVGP is more uncertain on the decision boundary. All this motivates the use of DGPs instead of SVGPs for problems that require the capture of complex patterns.

References

[1] J. Wassermann, *IASPEI New Manual Seismological Observatory Practice*, vol. 1. Potsdam, Germany: Volcano Seismology, 2002, ch. 13, p. 42.
[2] B. Chouet, "Volcano seismology," *Pure Appl. Geophysics*, vol. 160, no. 3, pp. 739–788, Mar. 2003, doi: 10.1007/PL00012556.
[3] D. J. Lary, A. H. Alavi, A. H. Gandomi, and A. L. Walker, "Machine learning in geosciences and remote sensing," *Geosci. Frontiers*, vol. 7, no. 1, pp. 3–10, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1674987115000821
[4] Q. Kong, D. T. Trugman, Z. E. Ross, M. J. Bianco, B. J. Meade, and P. Gerstoft, "Machine learning in seismology: Turning data into insights," *Seismological Res. Lett.*, vol. 90, no. 1, pp. 3–14, Jan. 2019, doi: 10.1785/0220180259.
[5] K. J. Bergen, P. A. Johnson, M. V. de Hoop, and G. C. Beroza, "Machine learning for data-driven discovery in solid Earth geoscience," *Science*, vol. 363, no. 6433, Mar. 2019, Art. no. eaau0323. [Online]. Available: https://science.sciencemag.org/content/363/6433/eaau0323

[1] Results become stable after a few samples. Here, S was set to 100.

[2] https://github.com/ICL-SML/Doubly-Stochastic-DGP

[6] E. Del Pezzo, "Discrimination of earthquakes and underwater explosions using neural networks," *Bull. Seismological Soc. Amer.*, vol. 93, no. 1, pp. 215–223, Feb. 2003, doi: 10.1785/0120020005.

[7] Q. Kong, R. M. Allen, and L. Schreier, "MyShake: Initial observations from a global smartphone seismic network," *Geophys. Res. Lett.*, vol. 43, no. 18, pp. 9588–9594, Sep. 2016. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016GL070955

[8] M. Curilem *et al.*, "Feature analysis for the classification of volcanic seismic events using support vector machines," in *Nature-Inspired Computation and Machine Learning* (Lecture Notes in Computer Science), vol. 8857. Cham, Switzerland: Springer, 2014, pp. 160–171.

[9] F. Giacco, A. M. Esposito, S. Scarpetta, F. Giudicepietro, and M. Marinaro, "Support vector machines and MLP for automatic classification of seismic signals at Stromboli volcano," in *Proc. Conf. Neural Nets WIRN, 19th Italian Workshop Neural Nets*, Vietri Sul Mare, Italy, May 2009, pp. 116–123.

[10] G. Curilem, J. Vergara, G. Fuentealba, G. Acuña, and M. Chacón, "Classification of seismic signals at villarrica volcano (Chile) using neural networks and genetic algorithms," *J. Volcanology Geothermal Res.*, vol. 180, no. 1, pp. 1–8, Feb. 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0377027308006355

[11] M. C. Benitez *et al.*, "Continuous HMM-based seismic-event classification at deception island, antarctica," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 1, pp. 138–146, Jan. 2007.

[12] M. Beyreuther and J. Wassermann, "Continuous earthquake detection and classification using discrete hidden Markov models," *Geophys. J. Int.*, vol. 175, no. 3, pp. 1055–1066, Dec. 2008. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-246X.2008.03921.x

[13] P. B. Dawson, M. C. Benítez, J. B. Lowenstern, and B. A. Chouet, "Identifying bubble collapse in a hydrothermal system using hidden Markov models," *Geophys. Res. Lett.*, vol. 39, no. 1, Jan. 2012. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/action/showCitFormats?doi=10.1029%2F2011GL049901

[14] G. Cortés, L. García, I. Álvarez, C. Benítez, Á. de la Torre, and J. Ibáñez, "Parallel system architecture (PSA): An efficient approach for automatic recognition of volcano-seismic events," *J. Volcanology Geothermal Res.*, vol. 271, pp. 1–10, Feb. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0377027313002229

[15] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006, doi: 10.1162/neco.2006.18.7.1527.

[16] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[17] Z. E. Ross, M.-A. Meier, and E. Hauksson, "P Wave arrival picking and first-motion polarity determination with deep learning," *J. Geophys. Res., Solid Earth*, vol. 123, no. 6, pp. 5120–5129, Jun. 2018. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2017JB015251

[18] Y. Wu, Y. Lin, Z. Zhou, D. C. Bolton, J. Liu, and P. Johnson, "DeepDetect: A cascaded region-based densely connected network for seismic event detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 62–75, Jan. 2019.

[19] S. M. Mousavi, W. Zhu, W. Ellsworth, and G. Beroza, "Unsupervised clustering of seismic signals using deep convolutional autoencoders," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 11, pp. 1693–1697, Nov. 2019.

[20] M. Titos, A. Bueno, L. Garcia, and C. Benitez, "A deep neural networks approach to automatic recognition systems for volcano-seismic events," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 5, pp. 1533–1544, May 2018.

[21] M. Titos, A. Bueno, L. Garcia, M. C. Benitez, and J. Ibanez, "Detection and classification of continuous volcano-seismic signals with recurrent neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 1936–1948, Apr. 2019.

[22] M. Titos, A. Bueno, L. García, C. Benítez, and J. C. Segura, "Classification of isolated volcano-seismic events based on inductive transfer learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 5, pp. 869–873, Aug. 2019.

[23] A. Bueno, C. Benítez, S. De Angelis, A. Díaz Moreno, and J. M. Ibáñez, "Volcano-seismic transfer learning and uncertainty quantification with Bayesian neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 2, pp. 892–902, Oct. 2019.

[24] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning* (Adaptive Computation and Machine Learning). Cambridge, MA, USA: MIT Press, 2006.

[25] J. Lee, J. Sohl-dickstein, J. Pennington, R. Novak, S. Schoenholz, and Y. Bahri, "Deep neural networks as Gaussian processes," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–17. [Online]. Available: https://openreview.net/forum?id=B1EA-M-0Z

[26] N. Lawrence. (2016). *Deep Learning, Pachinko, and James Watt: Efficiency is the Driver of Uncertainty*. [Online]. Available: http://inverseprobability.com/2016/03/04/deep-learning-and-uncertainty

[27] A. Kendall. (2017). *Deep Learning is Not Good Enough, We Need Bayesian Deep Learning for Safe Ai*. [Online]. Available: https://alexgkendall.com/computer_vision/bayesian_deep_learning_for_safe_ai

[28] Y. Gal, "Uncertainty in deep learning," Ph.D. dissertation, Univ. Cambridge, Cambridge, U.K., 2016.

[29] A. Damianou and N. Lawrence, "Deep Gaussian processes," *Artif. Intell. Statist.*, vol. 3, pp. 207–215, 2013.

[30] H. Salimbeni and M. Deisenroth, "Doubly stochastic variational inference for deep Gaussian processes," in *Advances in Neural Information Processing Systems 30*. Red Hook, NY, USA: Curran Associates, 2017, pp. 4588–4599.

[31] D. Moore and S. Russell, "Signal-based Bayesian seismic monitoring," in *Proc. 20th Int. Conf. Artif. Intell. Statist.*, vol. 54, A. Singh and J. Zhu, Eds. Fort Lauderdale, FL, USA: PMLR, Apr. 2017, pp. 1293–1301. [Online]. Available: http://proceedings.mlr.press/v54/moore17a.html

[32] D. A. Moore and S. J. Russell, "Gaussian process random fields," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, vol. 2. Cambridge, MA, USA: MIT Press, 2015, pp. 3357–3365.

[33] M. Noori, H. Hassani, A. Javaherian, H. Amindavar, and S. Torabi, "Automatic fault detection in seismic data using Gaussian process regression," *J. Appl. Geophys.*, vol. 163, pp. 117–131, Apr. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0926985118301964

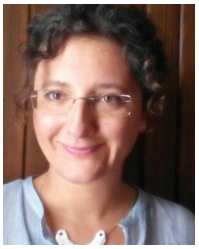
[34] J. Hensman, A. G. G. de Matthews, and Z. Ghahramani, "Scalable variational Gaussian process classification," in *Proc. 18th Int. Conf. Artif. Intell. Statist. (AISTATS)*, San Diego, CA, USA, May 2015, pp. 351–360. [Online]. Available: http://jmlr.org/proceedings/papers/v38/hensman15.html

[35] M. Palo *et al.*, "Analysis of the seismic wavefield properties of volcanic explosions at Volcán de colima, México: Insights into the source mechanism," *Geophys. J. Int.*, vol. 177, no. 3, pp. 1383–1398, Jun. 2009, doi: 10.1111/j.1365-246X.2009.04134.x.

[36] G. Cortes *et al.*, "Evaluating robustness of a HMM-based classification system of volcano-seismic events at colima and popocatepetl volcanoes," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, vol. 2, Jul. 2009, pp. II-1012–II-1015.

Miguel López-Pérez received the B.Sc. degree in mathematics and the M.S. degree in data science and computer engineering from the Universidad de Granada, Granada, Spain, in 2017 and 2018, respectively, where he is pursuing the Ph.D. degree under the supervision of Prof. Molina and Prof. Katsaggelos.

He is a member of the Visual Information Processing Group, Department of Computer Science and Artificial Intelligence, Universidad de Granada. His research interests focus on the use of Bayesian modeling, especially the Gaussian processes and their application to image processing, computer vision, and classification problems, working usually with medical imaging problems.



Luz García received the M.Sc. degree in telecommunication engineering from the Polytechnic University of Madrid, Madrid, Spain, in 2000, and the Ph.D. degree from the University of Granada, Granada, Spain, in 2008.

After working as a Support Engineer for communication networks at Ericsson-Spain, Madrid, for five years, she joined a European research project at the University of Granada. She has been with the Department of Signal Theory, Telematics, and Communications, University of Granada, since 2005, where she is an Assistant Professor. Her research interests are signal processing, pattern recognition, distributed acoustic sensing, and machine learning.

Carmen Benítez was a Visiting Researcher with the International Computer Science Institute, Berkeley, CA, USA, in 2000, and a Visiting Researcher with United Stated Geological Survey, Menlo Park, CA, USA, in 2009. From 2015 to 2018, she was the Head of the Department of Signal Theory, Telematics, and Communications, University of Granada, Granada, Spain, where she is a Full Professor of signal theory and communications. Her research interests include signal processing, geophysical signal processing, speech processing, machine learning, and pattern recognition. She has been involved in research projects funded by the Spanish Government, the European Union, and other third countries focused on the abovementioned research lines.

Rafael Molina (Senior Member, IEEE) received the M.Sc. degree in mathematics (statistics) and the Ph.D. degree in optimal design in linear models from the University of Granada, Granada, Spain, in 1979 and 1983, respectively.

He was the Dean of the Computer Engineering School, University of Granada, from 1992 to 2002, where he became a Professor of computer science and artificial intelligence in 2000. He was the Head of the Computer Science and Artificial Intelligence Department, University of Granada, from 2005 to 2007. He has coauthored an article that received the runner-up prize at reception for early stage researchers at the House of Commons in 2007. He has coauthored an awarded Best Student Paper at the IEEE International Conference on Image Processing in 2007, the ISPA Best Paper in 2009, and the EUSIPCO 2013 Best Student Paper. His research interest focuses mainly on using Bayesian modeling and inference in image restoration (applications to astronomy and medicine), super-resolution of images and video, blind deconvolution, computational photography, source recovery in medicine, compressive sensing, low-rank matrix decomposition, active learning, fusion, supervised learning, and crowdsourcing.

Dr. Molina has served as an Associate Editor for *Applied Signal Processing* from 2005 to 2007 and the IEEE TRANSACTIONS ON IMAGE PROCESSING from 2010 to 2014. He has been serving as an Area Editor for *Digital Signal Processing* since 2011.